DOCUMENT RESUME

ED 462 439 TM 033 703

AUTHOR Thornton, Artist
TITLE A Primer on the 2- and 3-Parameter Item Response Theory Models.
PUB DATE 2002-02-01
NOTE 17p.; Paper presented at the Annual Meeting of the College of Education, University of North Texas, Educational Research Exchange (2nd, Denton, TX, February 1, 2002).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Item Response Theory; Mathematical Models
IDENTIFIERS *Three Parameter Model; *Two Parameter Model

ABSTRACT

Item response theory (IRT) is a useful and effective tool for item response measurement if used in the proper context. This paper discusses the sets of assumptions under which responses can be modeled while exploring the framework of the IRT models relative to response testing. The one parameter model, or one parameter logistic model, is perhaps the simplest of IRT models. It presumes that only a single item parameter is necessary to represent the item response procedure. This parameter distinction is termed difficulty and given the symbol, beta. All unidimensional IRT models operate on the belief that a single fundamental latent construct (theta) is the chief contributory determinant of the experimental responses to each test's items. In the two-parameter model, the discrimination parameter, or the Greek symbol of alpha, is added. This allows the item characteristic curves (ICCs) for different items to exhibit different slopes. This discrimination parameter allows the modeling of the fact that some items have powerful (or feeble) associations to the fundamental construct being evaluated (theta). The three-parameter model adds one more parameter to the two-parameter model to reveal the reality that the lower asymptote of the ICC in accounting for guessing may well require the acceptance of nonzero values for their effective minimum values. The paper reviews some studies involving the use of IRT and discusses the ways in which IRT benefits research and testing. (Contains 24 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 462 439

A Primer on the 2- and 3- Parameter Item Response Theory Models

Artist Thornton

University of North Texas

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

A Thornton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033703

Paper presented at the annual meeting of the Educational Research Exchange, University of North Texas, February 1, 200, Denton, TX.

BEST COPY AVAILABLE

Item Response Theory

Introduction

Because of the growing debate between the effectiveness of Item Response Theory and Classical Testing Theory, it becomes more imperative for the proponents of IRT to demonstrate its usefulness in education research. Originally, IRT methods were used principally with standardized achievement and aptitude tests organized with multiple guess items dichotomously scored (Harvey Hammer 1999). While item response theory methods have been reality for more than half a century, only of late have they begun to achieve extensive regard in psychological assessment (Harvey and Hammer 1999). The uniqueness of this model is often associated with the unidimensional measure in testing. According to Harvey and Hammer (1999), "The IRT based approach to test development has the advantage of allowing the test developer to easily determine the effect of adding, or deleting, a given test item or set of test items by examining the test information function and/or the standard error function for an item pool" (p.367). However, this same uniqueness has been the argument for ineffectiveness. Again, Harvey and Hammer (1999) suggest that it is important to understand IRT limitations in terms of usefulness and effectiveness. Researchers are asked to remember that when measuring, one is fitting to a mathematical model with certain

assumptions and limitations. Harvey and Hammer (1999) argue that there does not exist any guarantee that the models involved with a given IRT approach, whether it being one parameter versus two parameter versus three parameter, will offer a sufficient data fit.

As mentioned before, IRT models deal primarily with standardized test with a dichotomous set up (ie. . . items scored right or wrong, true false,). According to Ayala and Bolesta (1999), "IRT is used in state testing programs, such as the Maryland State Department of Education High School Functional Assessment program and in municipal programs such as the Portland school District, for test equating" (p. 3). It is significant to underscore that the IRT models for dichotomous items are not limited to two alternative multiple choice arrangement, meaning they can be applied to multiple choice items that have any preferred number of response options and even to non-multiple choice items (Harvey & Hammer 1999). Essentially, the key prerequisite is that each person's item response has the capacity to be scored to manufacture a dichotomy. Additionally, IRT models deal with a poloychotomous or poloytomous fashion. According to Ayaloa and Bolesta (1999), "Most IRT work has been based on dichotomous models, however, not all examinee-item interaction can be modeled by a dichotomous model. For example, to capture the information in a Likert item or to assign credit for a partially correct answer requires a polytomous model" (p. 3). Polytomous models contain more item parameters than the dichotomous model so larger

samples are required. For example, Ayala and Bolesta (1999) note "larger ratio of examinees to item parameters was needed for Master's (1982) partial credit model to produce stable item and trait parameter estimates, regardless of the number of categories" (p. 4).

Item Response Theory is a useful and effective tool for item response measure if used in the proper context. According to Edward Hak-Sing (2001), "For more than a decade, the Item Response Theory (IRT) has provided a framework under which dichotomous and poloytomous responses to items can be modeled under a specific set of assumptions" (p. 109). This paper will discuss these specific sets of assumptions while exploring the framework of the IRT models relative to response testing.

One Parameter Model

The one parameter model or one parameter logistic model is perhaps the simplest of IRT models. As its title entails, it presumes that only a single item parameter is necessary to represent the item response procedure (Harvey & Hammer 1999). This parameter distinction is termed difficulty or given the symbol *b*. According to Harvey and Hammer (1999), "operationally, it [one parameter] is defined as the score on theta that is associated with 50% likelihood of a correct/endorsed item response" (p. 359). All uni-dimensional IRT models impart the belief that a single fundamental latent construct [theta] is the chief contributory

determinant of the experimental responses to each test's items (Harvey & Hammer 1999). In a study conducted by Prieto, Roset, and Badia, entitled *Rash Measurement in the Assessment of Growth Hormone Deficiency in Adult Patients* (2001), a sample of 356 repeated adult patients with untreated GHD was incorporated in the study. Patients answered the survey at 12 months apart. Responses were evaluated following the dichotomous logistic response model. Parameter approximates, model-data fit and separation statistics were calculated. The invariance of the item parameters across time was tested in the follow-up. Rasch results were furthermore employed to determine score differences through the computation of the Reliable Change Index (p. 49). One disadvantage to the 1-parameter model is its postulation that all items in the test share the same shaped ICC's; while this might be realistic in an item group that was quite abnormal in many practical measurement conditions (Harvey & Hammer 1999). According to Junker & Sijtsma (2000), "The monotonicity of item response functions is a central feature of most parametric and nonparametric item response models" (p. 65). Monotonicity permits items to be understood as calculating a trait, and it permits for a common speculation of nonparametric deduction traits (Junker & Sijtsma 2000).

Two Parameter Model

In the two parameter model, the discrimination parameter or *a* is added. This allows the ICC's for different items to exhibit different slopes. The discrimination parameter allows us to model the fact that some items have powerful (or feeble) associations to the fundamental construct being evaluated (theta); superior values denote firm associations (Harvey and Hammer 1999). According to Harvey & Hammer (1999), "The *a* parameter is very important in IRT due to the fact that it directly determines the amount of information provided by an item: Items with higher *a* parameters provide more information regarding theta, all other factors being equal" (p. 361). Rogers and Ndalichako (2000), illustrate the 2-parameter item response while performing an analysis of 1232 high school seniors. In their article, *Number-Right, Item-Response, and Finite-State Scoring: Robustness With Respect To Lack of Equally Classifiable Options And Item Option Independence* (2000), their analysis concluded that the number of right and 1 and 2 parameter methods were equally sensitive to the presence of absurd option of stem option connections and pairs of similar or opposite options (5).

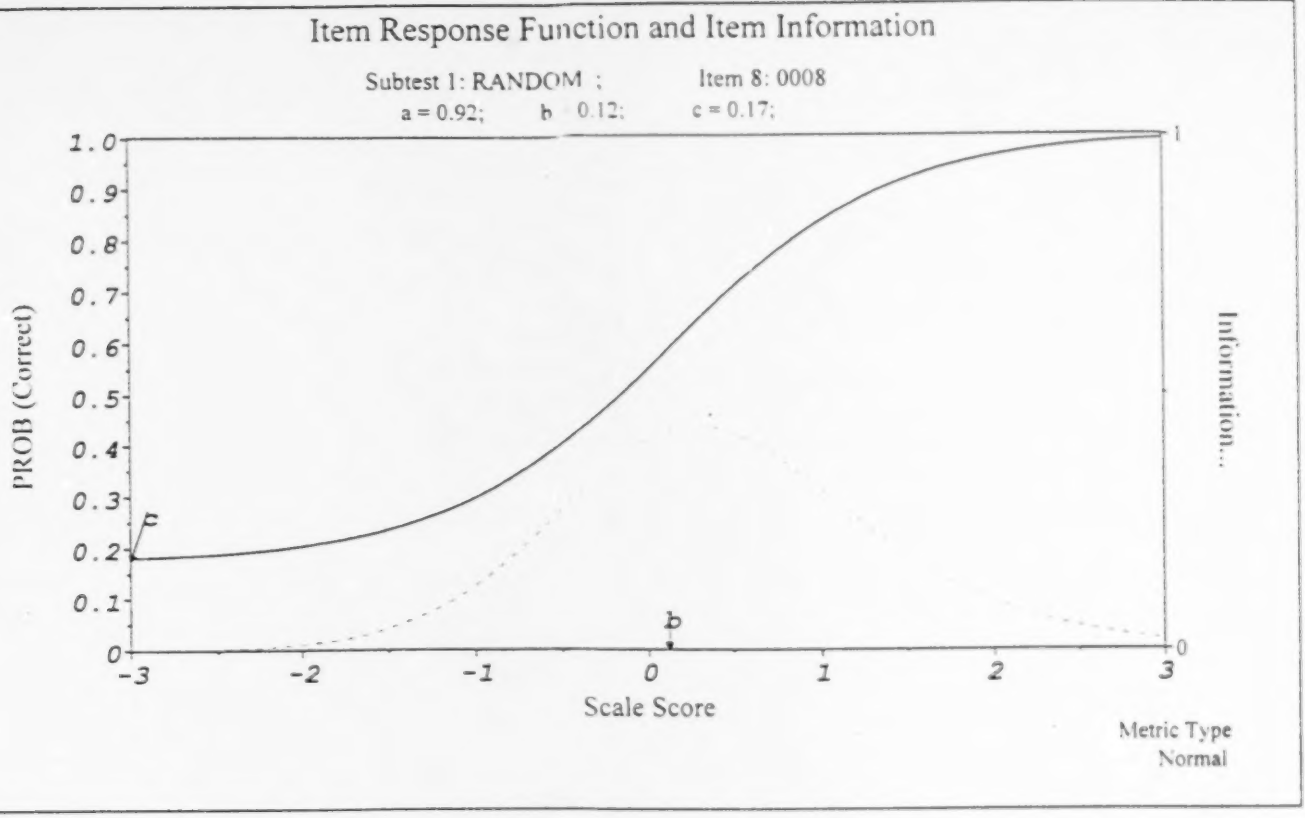
Three Parameter Model

Even though the 2 parameter model deals with one of the most grave assessments of the Rasch model ie., the postulation that all test items are alike with regard to their discriminating power), it does not

address another likely significant fact that may be different across items. The 3 parameter model adds one more parameter (c) to the two parameter model to reveal the reality that the lower asymptote of the ICC in accounting for guessing, may well require the acceptance of nonzero values for their effective minimum values (Harvey & Hammer 1999). In a study done by Hoskens and Boeck (2001), one can observe the function of a 3 parameter model. They state:

> The framework for modeling componential data using item response theory models for polytomous items is presented. This framework models response accuracies on complex cognitive tasks, which are decomposed in terms of more basic elements, such as knowledge structures, cognitive processes, and strategies" (19).

The following is a graph illustrating the Three- parameter model in Item Response Theory. It is readily observed how the one and two parameters compliment the third parameter within this model. The beta or difficulty level is .12, while the guessing parameter or third parameter is set at .17. One can observe that the discrimination parameter a is at .92 level, which illustrates that this item as compared to others would be good for testing.


Item Response Function and Item Information
Subtest 1: RANDOM ;
Item 8: 0008
a = 0.92;
b = 0.12;
c = 0.17;
PROB (Correct)
Information...
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
0
c
b
-3
-2
-1
0
1
2
3
Scale Score
Metric Type
Normal

Articles Detailing Studies Involving IRT

There exist several studies involving the use of Item Response Theory. A study by Bickel, Buyske, Chang, and Ying (2001) states, "An important assumption in IRT model-based adaptive testing is that matching difficulty levels of test items with an examinee's ability makes a test more efficient" (p. 69). In their article, the premise of their study deals with "when adding an item to a test, the improvement of accuracy is an increasing function of the item information, this assumption amounts to claiming that the item information is maximized when its difficulty level matches the examinee's ability" (p. 69). In another study, Ogasawara (2000) examines how asymptotic standard errors of item response theory are derived. According to Ogasawara (2000), "Two variations of the item and test response function methods and SEs of their parameter estimates are presented that use logit transformations of the item response functions" (p. 53). Ogasawara concludes that numerical examples show similarities between the small size SEs of the item and the small size test response function methods are actually smaller than those of other methods (p. 53). Another article examines how Rasch models differ by observing the contents of the Rasch measurement partial credit model and comparing it to other Rasch models. Bode (1998) discusses "The calibration of instruments with increasingly complex items is described, starting with dichotomous items

and moving on the poloychotomous items using a single rating scale, and mixed polychotomous items using multiple rating scales, and instruments in which each item has its own rating scale (p. 78).

Another study, which outlines the use of the discrimination and guessing parameter entitled, *Optimal Item Discrimination and Maximum Information for Logistic IRT Models* was done by Veerkamp and Berger (1999). According to their article, "This study derives discrimination parameter values, as functions of the guessing parameter and distances between person parameters and item difficulty, that yield maximum information for the three-parameter logistic item response theory model" (p. 31).

Reise (2000) discusses in his article, *Using multilevel Regression to Evaluate Person-Fit IRT Models*, "how multilevel logistic regression can be used to assess the consistency of an individuals response pattern with an item response theory measurement model" (p. 543). In Kamata's (2001) article, Item Analysis by the Hierarchical Generalized Linear Model, the hierarchical generalized model is presented as an explicit two-level formulation of multilevel item response model" (p. 79). Discussion is provided that explains how the HGLM model is equivalent to the Rasch model as well as an examination of how the characteristics of the HGLM model can be expressed as a latent regression model (pp. 79-93).

Fernando, Lorenzo and Molina (2001) discuss how the item response theory model of response stability is developed based on the local independence principle. In their article, *An Item Response Theory*

Analysis of Response Stability in Personality Measurement (2001), they examine how "the model predicts response changes under repeated administration of the same instrument using item and examinee parameter estimates as predictors" (p. 3). Muraki, Hombo and Lee (2000), discuss *Equating and Linking of Performance Assessments*, and offer an overview of linking methods applied to performance assessments. "Major issues and recent developments in linking performances are discussed. Three common linking designs (single group, randomly equivalent groups, and nonequivalent groups with anchor items) are compared (p. 325). Segall's (2001) article, *General Ability Measurement: An Application Of Multidimensional Item Response Theory*, discusses ways to improve measurement accuracy. "One method provides a multidimensional item response theory estimate obtained from conventional administration of multiple choice test items, while the other method chooses items adaptively to maximize the precision of the general ability" (p. 79). Schulz, Kolen, and Nicewander (1999), explore a new procedure for defining achievement tests in their article, *A Rationale for Defining Achievement Levels Using IRT-Estimated Domain Scores*. "This procedure assigns examinees to levels of achievement when the levels are represented by separate pools of multiple choice items. Items were assigned to levels on the basis of their content and hierarchically defined level descriptions" (p. 347). Wolfe's (2000), *Equating and Item Banking with the Rasch Model*, discusses the Rasch measurement procedures for

equating multiple test forms. "The procedures entail selecting an appropriate data collection design, estimating parameters, transforming the parameters from multiple forms to a common scale and evaluating the quality of the linkage between these forms" (p. 409).

Conclusion

Item Response Theory is historically the most widely used form of item analysis (Harvey & Hammer 1999). Only until recently has it gained popularity within educational research and psychological measurement. IRT has been widely accepted in standardized aptitude testing. According to Harvey & Hammer (1999), "one very practical reason for this belated popularity is the fact IRT techniques tend to be far more computationally demanding than methods of test construction and scoring that are based on classical test theory" (p. 353). Item Response Theory offers quality item response analysis that benefit researchers and psychologists alike. "IRT benefits research and testing by the fact that it provides a much more detailed view of item-level and test level functioning. It can be adapted to many different kinds of tests; the score estimation process is more precise, allowing simultaneous consideration of both the number of right/endorsed items as well as the properties (difficulty, discrimination) of each item" (Harvey & Hammer 1999).

References

Ayala, R.J., Bolesta, Monica. (1999) Item parameter recovery for the nominal response model. *Applied Psychological Measurement, 23* (1), 3-19.

Bickel, Peter, Buyske, Steven and Huahua Chang. (Mar 2001). On maximizing item information and matching difficulty with ability. *Psychometrika, 66* (1), 69-77.

Bode, Rita. (2001). Understanding rasch measurement: Partial credit model and pivot anchoring. *Journal of Applied Measurement, 2* (1), 78-95.

Childs, Ruth, Chen, Wen-Hung. (1999). Obtaining comparable item parameter estimates in MULTILOG and PARASCALE. *Applied Psychological Measurement, 23* (4), 371-79.

Ferrando, Pere, Lorenzo, Urbano and Gabriel Molina. (2001). An item response theory analysis of response stability in personality measurement. *Applied Psychological Measurement, 25* (1), 3-17.

Florin, Robert. (2000). Rasch analysis in measurement of physician work. *Journal of Outcome Measurement, 4* (2) 564-578.

Hak-sing, Edward. (2001). Testing for local dependency in dichotomous and poloytomous item response models. *Psychometrika, 66* (1), 109-132.

Harvey, Robert & Hammer, Allen. (1999). *Item Response Theory. Counseling Psychologist, 27* (3) 353-83.

Hombo, Catherine & Lee, Yong-Won. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24* (4), 325-337.

Hoskens, Matcteld & DeBoeck, Paul. (2001). Multidimensional componential item response theory models for polytomous Items. *Applied Psychological Measurement, 25* (1), 19-37.

Junker, Brian & Sijtsma, Klaas. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 2*4 (1), 65-81.

Kamata, Akihito. (2001). Item analysis by the hierarchical generalized linear model. *Journal Of Educational Measurement, 38* (1), 79-93.

Maydeu-Plivares. (1997). Handbook of modern item response theory. *Psychological Measurement, 21* (2), 185-87.

McDonald, Roderick. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24* (2), 99-114.

Mellenbergh, Gideon. (2001). The new rules of measurement: What every psychologist and educator should know. *Psychometrika, 66* (1), 153-155.

Ogasawara, Haruhiko. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement, 25* (1), 53-67.

Prieto, Luis, Roset, Montse, & Xavier Badia. (2001). Rasch measurement in the assessment of growth hormone deficiency in adult patients. *Journal of Applied Measurement, 2* (1), 48-64.

Reise, Steven. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35* (4) 543-68.

Roid, Gale. (2000). Use of rasch scaling in the measurement of cognitive development and growth. *Journal of Outcome Measurement, 4* (2) 579-94.

Rogers, Todd & Ndalichako, Joyce. (2000). Number-right, item-response, and finite-state scoring: Robustness with respect to lack of equally classifiable options and item option independence. *Educational and Psychological Measurement, 60* (1), 5-19.

Schultz, Matthew, Kolen, Michael, & W. Nicewander. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23* (4), 237-62.

Segall, Daniel. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika, 66* (1), 79-97.

Veerkamp, Wimn & Berger, Martin. (1999). Optimal item discrimination and maximum information for logistic IRT models. *Applied Psychological Measurement, 23* (1) 31-40.

Wolfe, Edward. (2000). Understanding rasch measurement: equating and item banking with the rasch model. *Journal of Applied Measurement, 1 (4)* 409-34.